Hodges-Lehmannovská optimálnosť testov pomerom vierohodností.

FRANTIŠEK RUBLÍK
ÚSTAV MERANIA SAV, BRATISLAVA

Nech sú pravdepodobnosti $\{\overline{P}_\gamma;\ \gamma \in \Xi\}$ definované na $(X, \mathcal{F})$ pomocou hustôt

$$f(x,\gamma) = \frac{\mathrm{d}\overline{P}_\gamma}{\mathrm{d}\nu}(x).$$

Predpokladajme, že $n_1^{(u)}, \ldots, n_u^{(q)}$, $u = 1, 2, \ldots$ sú rozsahy nezávislých výberov, a v tomto zápise

$$y(1, n_u^{(1)}) = (x_1^{(1)}, \ldots, x_{n_u^{(1)}}^{(1)}), \quad \ldots, \quad y(q, n_u^{(q)}) = (x_1^{(q)}, \ldots, x_{n_u^{(q)}}^{(q)})$$

označuje náhodné výbery z q štatistických populácií. Teda

$$x^{(u)} = (y(1, n_u^{(1)}), \ldots, y(q, n_u^{(q)})) \tag{1}$$

je združený výber, a v označení

$$\Theta = \Xi^q \tag{2}$$

je $\theta = (\theta_1, \ldots, \theta_q) \in \Theta$ parameter, popisujúci celkový stav tých q populácií, t.j. θ_j je parameter j-tej populácie. Cieľom je testovať hypotézu

$$\emptyset \neq \Omega_0 \subset \Theta.$$

Nech v označení

$$n_u = \sum_{j=1}^{q} n_u^{(j)}, \qquad p_u^{(j)} = n_u^{(j)}/n_u \tag{3}$$

platí, že

$$\lim_{u\to\infty} n_u = +\infty, \qquad \lim_{u\to\infty} p_u^{(j)} = p_j \in (0, 1), \quad j = 1, \ldots, q. \tag{4}$$

Ak je test φ_u nulovej hypotézy Ω_0 oproti $\Theta - \Omega_0$ založený na (1) a pre $\beta_u(\theta) = E_\theta(\varphi_u)$ platí, že asymptotická úroveň významnosti

$$\sup\{\lim_{u\to\infty} \beta_u(\theta^*);\ \theta^* \in \Omega_0\} = \alpha \in (0, 1), \tag{5}$$

tak pre každý parameter $\theta \in \Theta - \Omega_0$ je rýchlosť konvergencie pravdepodobnosti $1 - \beta_u(\theta)$ chyby druhého druhu ku nule ohraničená zdola, presnejšie

$$\liminf_{u\to\infty} \frac{1}{n_u} \log[1 - \beta_u(\theta)] \geq -K(\Omega_0, \theta, p), \tag{6}$$

kde $K(\Omega_0,\theta,p)$ Kulback-Leiblerovo informačné číslo; $K(\Omega_0,\theta,p)$ je definované ako

$$K(\Omega_0,\theta,p)=\inf\{K(\theta^*,\theta,p);\ \theta^*\in\Omega_0\}\,,$$

kde

$$K(\theta^*,\theta,p)=\sum_{j=1}^{q}p_jK(\theta_j^*,\theta_j)$$

a

$$K(\gamma^*,\gamma)=\int\log\left(\frac{f(x,\gamma^*)}{f(x,\gamma)}\right)f(x,\gamma^*)\,\mathrm{d}\nu(x)\,.$$

Dôkaz tejto nerovnosti možno nájsť napríklad v [4]. V zhode s ňou sa testy $\{\varphi_u\}$ nazývajú **Hodges-Lehmannovsky optimálne**, ak (4) a (5) implikujú, že

$$\lim_{u\to\infty}\frac{1}{n_u}\log[1-\beta_u(\theta)]=-K(\Omega_0,\theta,p)\,.$$

Budeme sa venovať Hodges-Lehmannovskej optimálnosti štatistiky pomeru vierohodností, ktorá je pre testovanie hypotézy Ω_0 definovaná vzorcom

$$T_u(x^{(u)})=\log\frac{L(x^{(u)},\Theta)}{L(x^{(u)},\Omega_0)}\,,$$

kde

$$L(x^{(u)},A)=\sup\{\prod_{j=1}^{q}\prod_{i=1}^{n_u^{(j)}}f(x_i^{(j)},\theta_j);\ (\theta_1,\ldots,\theta_q)\in A\}\,.$$

Ako je dobre známe, test Ω_0 oproti $\Theta-\Omega_0$ sa vykoná podľa pravidla

$$\varphi_u(x^{(u)})=\begin{cases}1 & T_u>t_u\,,\\ 0 & T_u\le t_u\,,\end{cases}$$

kde 1 označuje zamietnutie, 0 prijatie hypotézy, a t_u je zvolená kritická konštanta.

Jeden z prístupov ku tejto problematike možno nájsť v [3], kde sa predpokladá, že hustoty $\{f(x,\gamma);\ \gamma\in\Xi\}$ tvoria exponenciálnu triedu pravdepodobností, t.j. $\Xi\subset R^k$ a

$$f(x,\gamma)=\frac{\mathrm{d}\overline{P}_\gamma}{\mathrm{d}\nu}(x)=\mathrm{e}^{\gamma' x-C(\gamma)}\,.$$

Ak $\overline{x}_n\in\{\mathrm{E}\gamma(x);\ \gamma\in\Xi\}$, tak pre parameter $\hat{\theta}_n=\hat{\theta}_n(x_1,\ldots,x_n)\in\Xi$ určený rovnosťou

$$E_{\hat{\theta}_n}(y)=\overline{x}_n$$

po algebraických úpravách

$$\log L(x_1,\ldots,x_n,\gamma)=g_n(x_1,\ldots,x_n)-nK(\hat{\theta}_n,\gamma)$$

a $\hat{\theta}_n$ je OMV parametra γ. V [3] sa H-L optimálnosť štatistiky pomerom vierohodností dokazuje za rôznych predpokladov podľa toho, či sa jedná o výber z jednej, alebo z $q > 1$ štatistických populácií.

Jednovýberový prípad q=1

Označme $A_n = \{\gamma \in \Omega_0;\ K(\gamma, \Omega_0) \le \frac{t_n}{n}\}$. Ak

$$\overline{x}_n = E_\gamma(x)\,,\ K(\gamma,\theta) < K(A_n,\theta)$$

tak $\hat{\theta}_n = \gamma \notin A_n$ a teda štatistika pomeru vierohodností

$$T_n(x_1,\ldots,x_n) = nK(\hat{\theta}_n,\Omega_0) > \frac{t_n}{n}.$$

Preto

$$\begin{aligned}1-\beta_n(\theta) &= P_\theta(T_n(x_1,\ldots,x_n) \le t_n) \le \\ &\le P_\theta(\overline{x}_n \notin \{E_\gamma(x);\ K(\gamma,\theta) < K(A_n,\theta)\}).\end{aligned}$$

Podľa vety 3.2.1 v [2]

$$P_\theta(\overline{x}_n \notin \{E_\gamma(x);\ K(\gamma,\theta) < d_n\}) = (nd_n)^{\frac{k-2}{2}} \exp(-nd_n + O(1))$$

pre $n \to \infty$, ak existuje $\varepsilon > 0$ tak, že

$$\frac{\varepsilon}{n} \le d_n \le \min\{\frac{1}{\varepsilon}, K(\theta)\},$$

$$K(\theta) = \sup\{a;\ \{\gamma \in \Xi;\ K(\gamma,\theta) \le a\} \subset C_a \subset \Xi,\ \ C_a \text{ je kompaktná}\}.$$

V [3] sa preto postuluje prepoklad $K(\Omega_0,\theta) < K(\theta)$ a spojitosť funkcie $K(.,\Omega_0)$, čo umožňuje použiť túto vetu 3.2.1 a pomocou nej za predpokladu $t_n = o(n)$ dokázať, že

$$\limsup \frac{1}{n}\log(1-\beta_n(\theta) \le -K(\Omega_0,\theta).$$

Tieto predpoklady sú obmedzením na nulovú hypotézu a treba ich overovať pre jednotlivé hypotézy zvlášť.

Viacvýberový prípad $q > 1$

Analóg vety 3.2.1 z [2] pre viacvýberový prípad nie je známy, a preto sa pre tento prípad v [3] používa iný prístup, ako pre $q = 1$. Postuluje sa predpoklad ktorý je ekvivalentný tomu, že odhad maximálnej vierohodnosti $\hat{\theta}$ existuje s pravdepodobnosťou 1 pre n dostatočne veľké. Potom s pravdepodobnosťou 1

$$\log L(x^{(u)},\theta^*) = G^{(u)}(x^{(u)}) - n_u K(\hat{\theta}_{(u)},\theta^*,p_u)$$

čo znamená, že

$$P_\theta[\log \frac{L(x^{(u)},\Theta)}{L(x^{(u)},\Omega_0)} \le t_u] = P_\theta[n_u K(\hat{\theta}_{(u)},\Omega_0,p_u) \le t_u] =$$

$$= P_\theta[\hat{\theta}_{(u)} \in A_u]$$

kde $A_u = \{\theta^* \in \Xi^q; K(\theta^*,\Omega_0,p) \le \frac{t_u}{n_u}\}$. Pomocou Herrovej nerovnosti]

$$P_\theta[\hat{\theta}_{(u)} \in A_u] \le P_\theta[K(\hat{\theta}_{(u)},\theta,p_u) \ge K(A_u,\theta,p_u)]$$

sa potom už dá dostať ku situácii, na ktorú možno aplikovať tvrdenia o pravdepodobnosti veľkých odchýlok typu

$$P_\theta\Big[\log \frac{L(x^{(u)},\Theta)}{L(x^{(u)},\theta)} \ge t\Big] \le \exp[-t + H_u(t)],$$

a pomocou postulovanej spojitosti funkcie $K(.,\Omega_0,p)$ dokázať H-L optimálnosť testov pomerom vierohodností. Tento prístup má okrem postulovanej spojitosti $K(.,\Omega_0,p)$ aj tú nevýhodu, že vylučuje Poissonovo a multinomické rozdelenie (presnejšie, systém všetkých pravdepodobností na konečnej množine, vedúci k multinomickému ako k rozdeleniu OMV), lebo tieto po reparametrizácii poskytujú exponenciálne triedy, pre ktoré OMV neexistuje s kladnou pravdepodobnosťou.

Axiomatický prístup v [5] pre dokázanie H-L optimálnosti štatistiky pomeru vierohodností umožňuje vyhnúť sa týmto ťažkostiam. V [5] sa predpokladá, že trieda pravdepodobností je zadaná hustotami $\{f(x,\gamma);\ \gamma \in \Xi\}$ vzhľadom na mieru ν, a sú splnené nasledujúce predpoklady.

(A I) Existuje σ-kompaktný metrický priestor Ξ_1 taký že Ξ je hustý v Ξ_1 (t.j. $\overline{\Xi} = \Xi_1$), a pôvodný sytém $\{f(x,\gamma);\gamma \in \Xi\}$ môže byť rozšírený do systému $\{f(x,\gamma);\gamma \in \Xi_1\}$ hustôt vzhľadom na tú istú mieru ν (rozšírením rozumieme že $f(x,\gamma)$ sa zhoduje s pôvodnou hustotou ak $\gamma \in \Xi$). Navyše, rozšírený systém $\{\overline{P}_\gamma;\gamma \in \Xi_1\}$ pozostáva z navzájom rôznych pravdepodobností, Kullback-Leiblerovo informačné číslo $K(\theta^*,.)$ je spojité na Ξ_1 pre každý parameter $\theta^* \in \Xi_1$ (t.j. $\gamma_n^* \to \gamma^*$ implikuje $K(\theta^*,\gamma_n^*) \to K(\theta^*,\gamma^*)$), a okrem toho $f(x,\gamma)$ je spojitá na Ξ_1 pre každý prvok $x \in X$.

(A II) Funkcia $K(.,\theta)$ je konečná a spojitá na Ξ_1 pre všetky $\theta \in \Xi$.

(A III) Nech parametre $\{\theta_n^*\}_{n=1}^\infty$ patria do Ξ_1, $\gamma \in \Xi$ a

$$\limsup_{n\to\infty} K(\theta_n^*,\gamma) < +\infty.$$

(a) Ak $\theta \in \Xi$, tak $\limsup_{n\to\infty} K(\theta_n^*,\theta) < +\infty$.

(b) Postupnosť $\{\theta_n^*\}_{n=1}^\infty$ má v Ξ_1 hromadnú hodnotu, a ak parametre $\{\theta_n\}_{n=1}^\infty$ patriace do Ξ sú také že $\limsup_{n\to\infty} K(\theta_n^*,\theta_n) < +\infty$, tak aj postupnosť $\{\theta_n\}_{n=1}^\infty$ má v Ξ_1 hromadnú hodnotu.

(c) Nech $\lim_{n\to\infty} \theta_n^* = \theta^*$. Ak $\{\theta_n\}_{n=1}^{\infty}$ je postupnosť parametrov zo Ξ a $\lim_{n\to\infty} \theta_n = \theta$, tak $\liminf_{n\to\infty} K(\theta_n^*, \theta_n) \geq K(\theta^*, \theta)$.

(A IV) Existujú merateľné množiny $A_n \subset X^n$, prirodzené číslo N a merateľné zobrazenia $\hat{\theta}_n : A_n \to \Xi_1$, $g_n : A_n \to R$ také, že pre všetky $\gamma \in \Xi$ a $n \geq N$

$$\overline{P}_\gamma(A_n) = 1 \tag{7}$$

a na množine A_n v označení $L(x_1, \ldots, x_n, \gamma) = \prod_{i=1}^n f(x_i, \gamma)$ platí rovnosť

$$\log L(x_1, \ldots, x_n, \gamma) = g_n(x_1, \ldots, x_n) - nK(\hat{\theta}_n, \gamma)\,. \tag{8}$$

(A V) V označeníz *(A 4)* a

$$p_u = (p_u^{(1)}, \ldots, p_u^{(q)})\,,$$

pre $\theta = (\theta_1, \ldots, \theta_q) \in \Xi^q$ nech

$$K\left(\hat{\theta}, \theta, p_u\right) = \sum_{j=1}^{q} p_u^{(j)} K\left(\hat{\theta}_{n_u^{(j)}}(y(j, n_u^{(j)})), \theta_j\right). \tag{9}$$

Ak platí (4), tak pre každý parameter $\theta \in \Xi^q$ a reálne číslo t

$$P_\theta[\,n_u K(\hat{\theta}, \theta, p_u) \geq t\,] \leq \exp[-t + h_u(t)] \tag{10}$$

kde funkcia h_u má tú vlastnosť, že

$$\limsup_{u\to\infty} \frac{h_u(t_u)}{n_u} = 0 \tag{11}$$

ak

$$\limsup_{u\to\infty} \frac{t_u}{n_u} < +\infty\,. \tag{12}$$

Tvrdenie o H-L optimálnosti má v [5] nasledujúci tvar.

Veta. Nech trieda pravdpodobností $\{\overline{P}_\gamma;\, \gamma \in \Xi\}$, určená hustotami $\{f(x, \gamma);\, \gamma \in \Xi\}$ vzhľadom na mieru ν je taká, že prepoklady *(A I) - (A V)* sú splnené. Nech

$$\emptyset \neq \Omega_0 \subset \Theta = \Xi^q\,,$$

$T_u = T_u(x^{(u)})$ je štatistika pomeru vierohodností pre testovanie tejto hypotézy, t.j. testujeme ju testami

$$\varphi_u(x^{(u)}) = \begin{cases} 1 & T_u > t_u\,, \\ 0 & T_u \leq t_u\,. \end{cases} \tag{13}$$

Ak z platnosti predpokladov

$$\lim_{u\to\infty} n_u = +\infty\,, \qquad \lim_{u\to\infty} p_u^{(j)} = p_j \in (0,1)\,, \quad j = 1,\ldots,q\,,$$

$$\sup\{\, \lim_{u\to\infty} \beta_u(\theta^*);\ \theta^* \in \Omega_0 \,\} = \alpha \in (0\,,1)\,,$$

vyplýva, že kritické konštanty t_u v (13) môžu byť vybrané tak, že

$$\limsup_{u\to\infty} t_u < +\infty\,, \tag{14}$$

tak testy (13) sú H-L optimálne.

Ak $\{\overline{P}_\gamma;\ \gamma \in \Xi\,\}$ sú regulárne normálne rozdelenia, alebo Poissonove rozdelenia, alebo ak $X = \{1,...,k\}$ je konečná množina a pravdepodobnosti na nej sú zadané pomocou

$$\Xi = \{\,(p_1,...,p_k)' \in R^k;\ \min_i p_i > 0,\ \sum_{i=1}^{k} p_i = 1\}\,, \qquad f(x,p) = p_x$$

tak pomocou predchádzajúcej vety možno dokázať toto tvrdenie.

Veta. Ak Ω_0 je neprázdna podmnožina množiny $\Theta = \Xi^q$ a ak T_u sú štatistiky pomeru vierohodností pre testovanie Ω_0 proti $\Theta - \Omega_0$, tak testy (13) založené na nich sú H-L optimálne pre testovanie Ω_0 oproti $\Theta - \Omega_0$.

CITÁCIE

[1] W. Hoeffding.: Asymptotically optimal tests for multinomial distributions. Annals of Mathematical Statistics 36(1965), 369–408.

[2] W. C. M. Kallenberg: Asymptotic optimality of likelihood ratio tests. Mathematisch Centre Tracts 77, Amsterdam 1978.

[3] S. Kourouklis: Hodges-Lehmann optimality of the likelihood ratio test in regular exponential families of distributions. Technical Report 73, The Pennsylvania State University, 1987.

[4] S. Kourouklis: Hodges-Lehmann efficacies of certain tests in multivariate analysis and regression analysis. The Canadian Journal of Statistics, 16(1988), 87-95.

[5] Rublík: On Hodges-Lehmann optimality of LR tests. Prijaté na publikáciu v časopise Kybernetika.